TOUR AND TRAVEL AGENCY

Data Analysis

AUGUST 29, 2021

SIDDHI PADEKAR

Table of Contents

Executive Summary

This Data set is of tour and travel agency, which deals in selling holiday packages. The dataset contains details of 872 employees of a company. Among these employees, some opted for the package and some did not. In this problem statement, we need to predict whether an employee will opt for the package or not. In addition, we need to find find out the important factors based on which the company will focus on particular employees to sell their packages.

Introduction

The purpose of this whole exercise is to explore the dataset. Do the exploratory data analysis. Explore the dataset, predict the holiday package on the bases of the details given in the dataset using logistic regression and LDA (linear discriminant analysis). Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC_AUC score for each model Final Model.

Data Dictionary:

Below is the brief description of data set

Variable Name	**Description**
Holiday_Package	Opted for Holiday Package yes/no?
Salary	Employee salary
age	Age in years
edu	Years of formal education
no_young_children	The number of young children (younger than 7 years)
no_older_children	Number of older children
foreign	foreigner Yes/No

Sample of the dataset:

	Unnamed: 0	Holliday_Package	Salary	age	educ	no_young_children	no_older_children	foreign
0	1	no	48412	30	8	1	1	no
1	2	yes	37207	45	8	0	1	no
2	3	no	58022	46	9	0	0	no
3	4	no	66503	31	11	2	0	no
4	5	no	66734	44	12	0	2	no

Above is sample details of data set showing first 5 rows.

Holiday Package is the target variable and all other are predictor variable

Exploratory Data Analysis

Data columns (total 8 columns):
Unnamed: 0 872 non-null int64
Holliday_Package 872 non-null object
Salary 872 non-null int64
age 872 non-null int64
educ 872 non-null int64
no_young_children 872 non-null int64
no_older_children 872 non-null int64
foreign 872 non-null object
dtypes: int64(6), object(2)

Shape of Data `(872, 8)`

1. The data set contains 872 row, 8 columns .
2. In the given data set there are 6 Integer type features , 2 Object type features.

Check for missing values in the dataset:

```
Unnamed: 0          0
Holliday_Package    0
Salary              0
age                 0
educ                0
no_young_children   0
no_older_children   0
foreign             0
dtype: int64
```

- There are No Nul values present in data set

Mean, Standard deviation , Min, Max

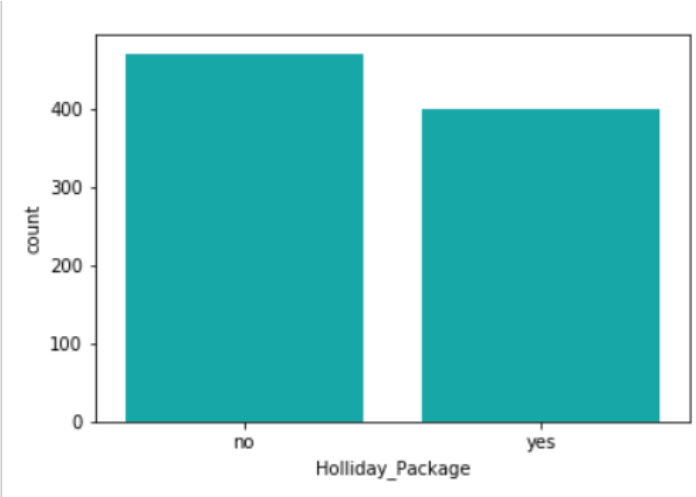
	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Unnamed: 0	872	NaN	NaN	NaN	436.5	251.869	1	218.75	436.5	654.25	872
Holliday_Package	872	2	no	471	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Salary	872	NaN	NaN	NaN	47729.2	23418.7	1322	35324	41903.5	53469.5	236961
age	872	NaN	NaN	NaN	39.9553	10.5517	20	32	39	48	62
educ	872	NaN	NaN	NaN	9.30734	3.03626	1	8	9	12	21
no_young_children	872	NaN	NaN	NaN	0.311927	0.61287	0	0	0	0	3
no_older_children	872	NaN	NaN	NaN	0.982798	1.08679	0	0	1	2	6
foreign	872	2	no	656	NaN	NaN	NaN	NaN	NaN	NaN	NaN

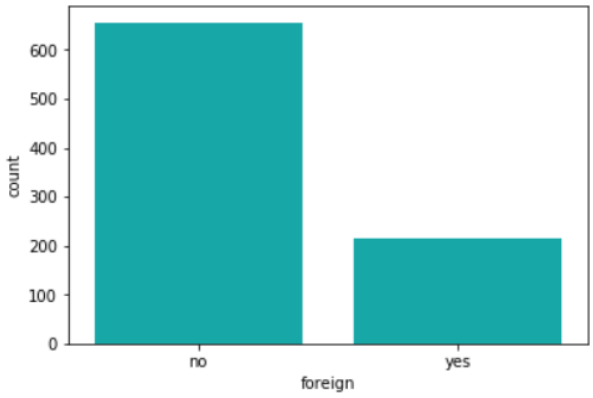
Categorising the object variable (Holiday Package , Foreign)

```
HOLLIDAY_PACKAGE :  2
yes    401
no     471
Name: Holliday_Package, dtype: int64
```

```
FOREIGN :  2
yes    216
no     656
Name: foreign, dtype: int64
```

- Holiday Package is categorised into 2 heads –Yes, No (Most employees has not opted for Holiday package)
- Foreign is categorised into 2 heads –Yes, No (There are less number of foreigners)





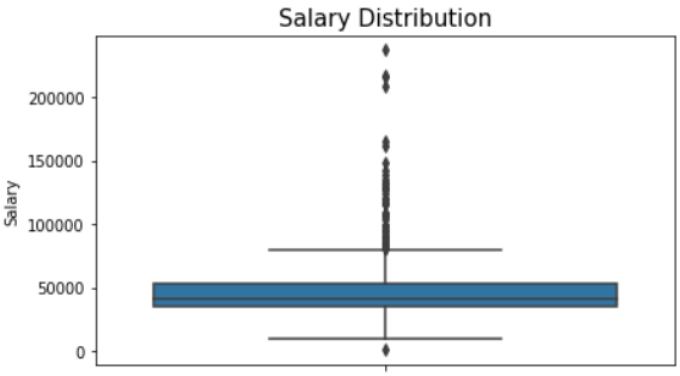
Check for Duplicate values and Removing Duplicate values in the dataset:

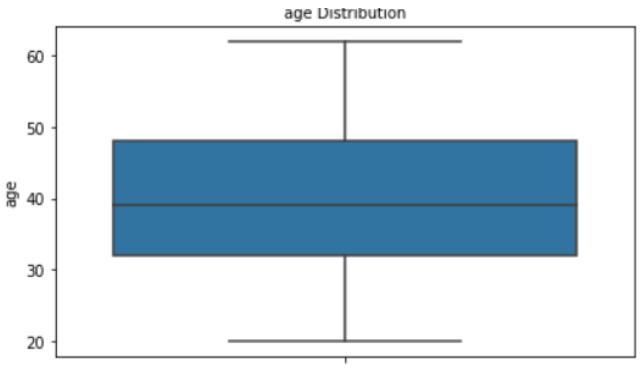
```
Number of duplicate rows = 0
```

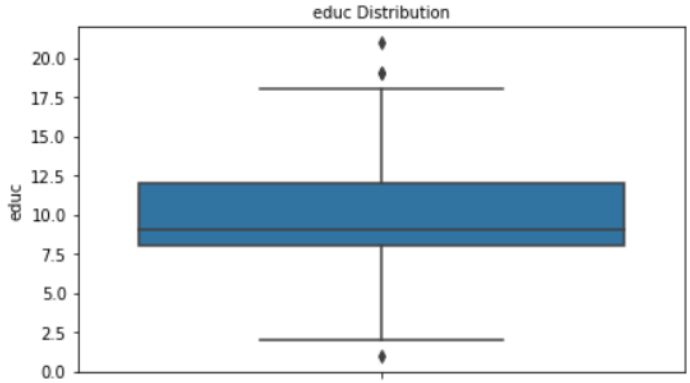
Note- we have dropped first column from the dataset ("Unnamed: 0") as this is only serial numbers

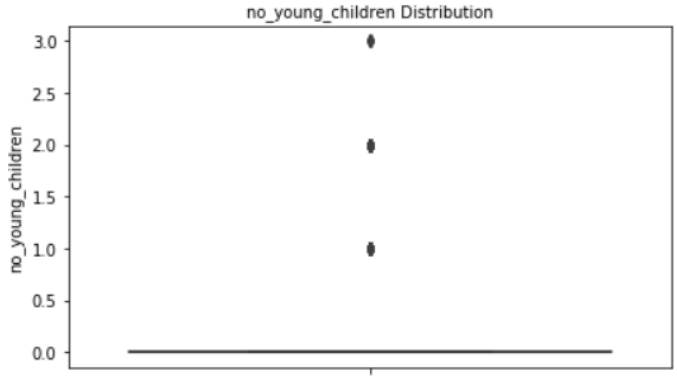
Outliers

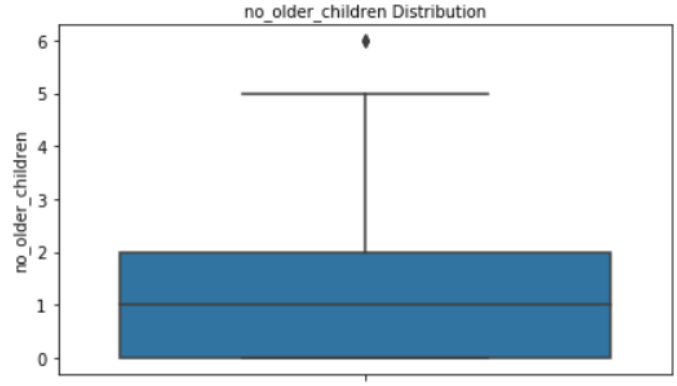
An outlier is an observation that lies an abnormal distance from other values in a random sample from a population. A value that "lies outside" (is much smaller or larger than) most of the other values in a set of data.







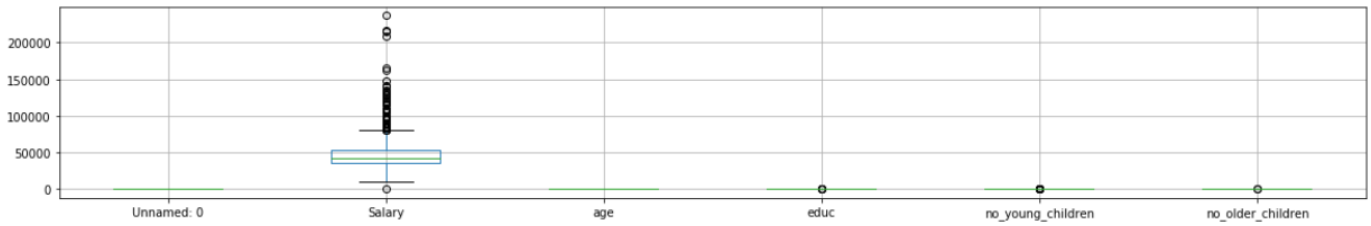




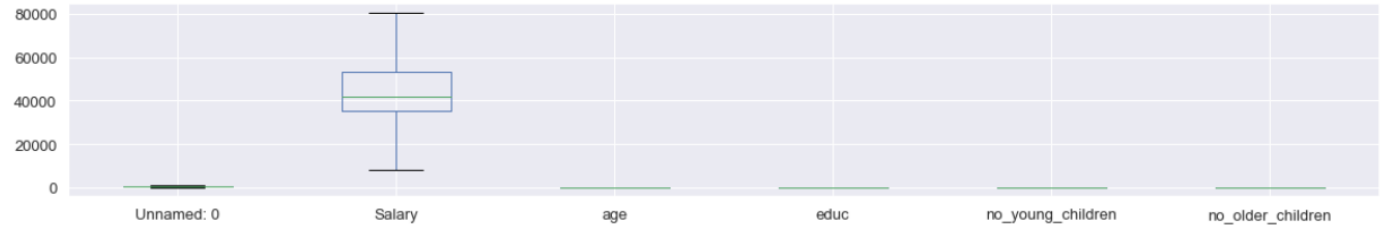
There are significant number of outliers in all the attributes

```
Unnamed: 0 --------- 0
Salary --------- 57
age --------- 0
educ --------- 4
no_young_children --------- 207
no_older_children --------- 2
```

Before removing Outliers

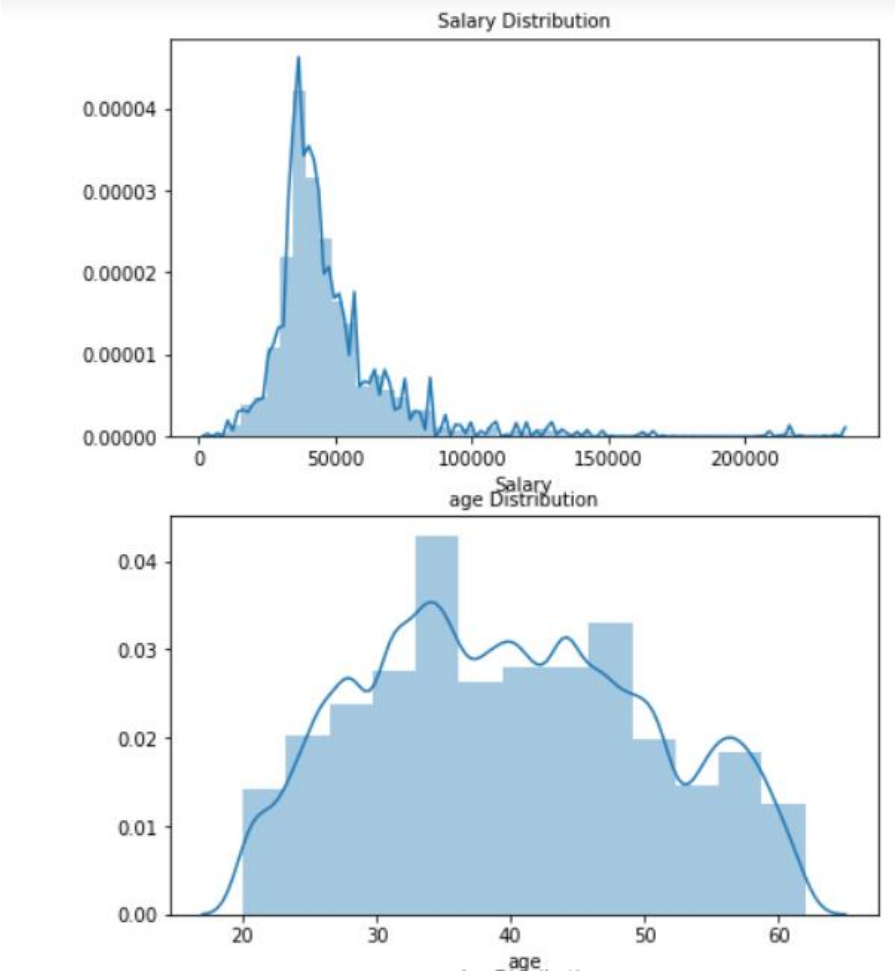


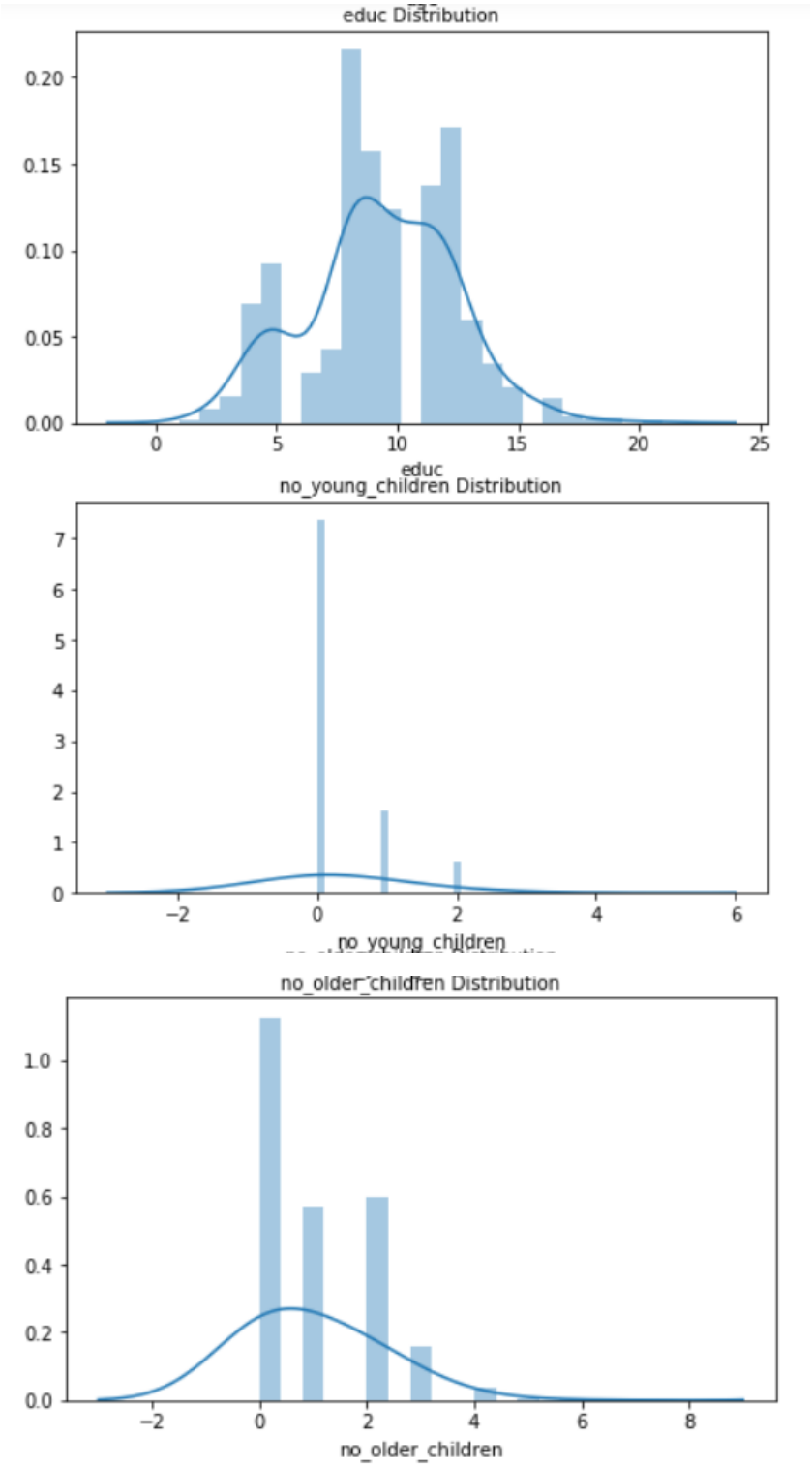
After removing outliers



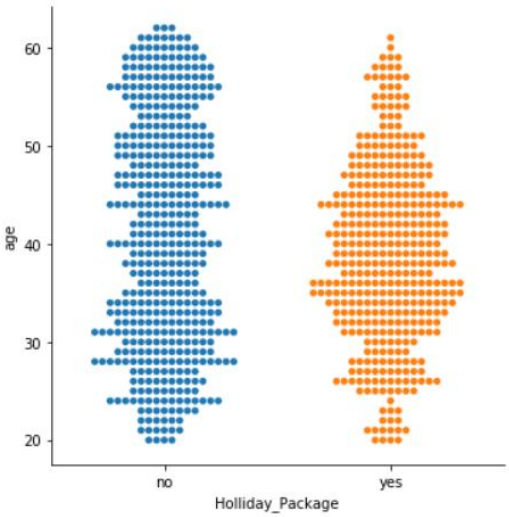
Univariate Analysis

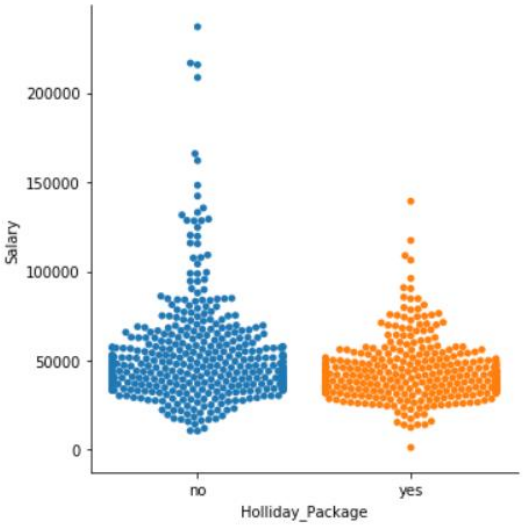
Histograms

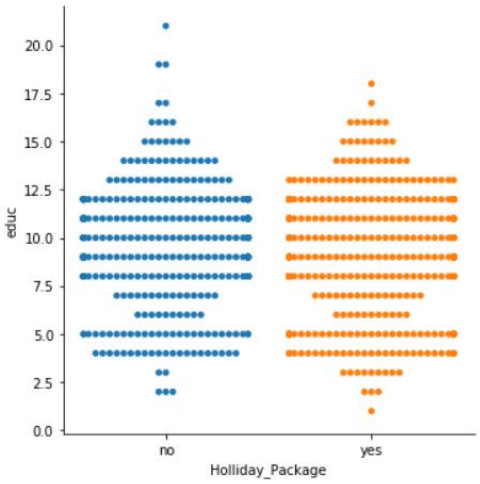


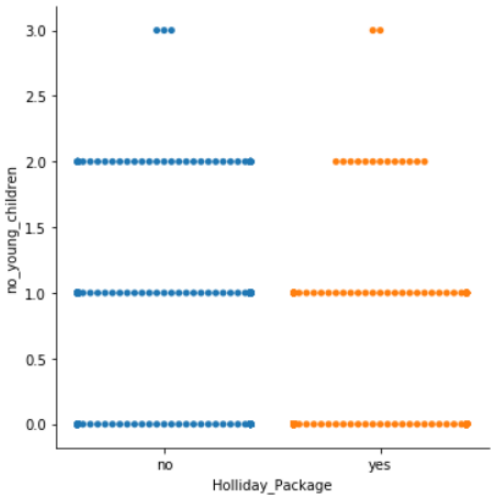


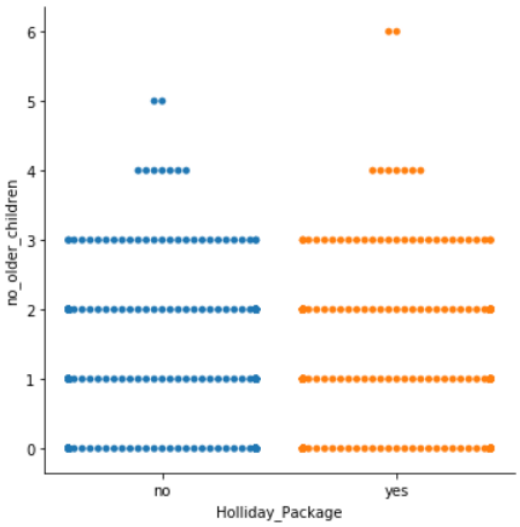
Bivariate Analysis

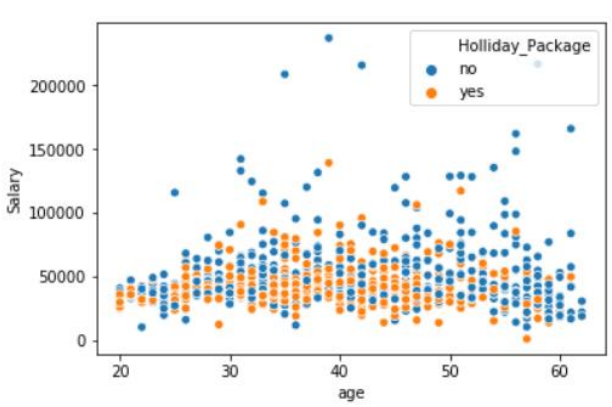

age
60
50
40
30
20
no
yes
Holliday_Package

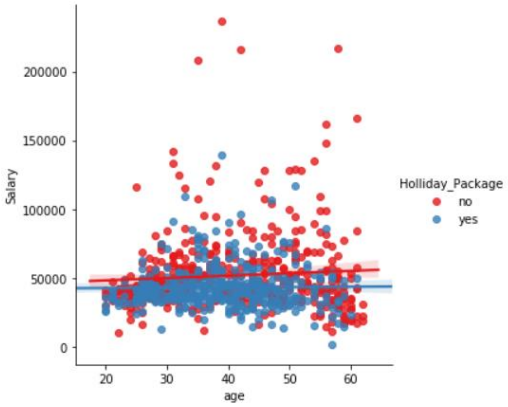

Salary
200000
150000
100000
50000
0
no
yes
Holliday_Package

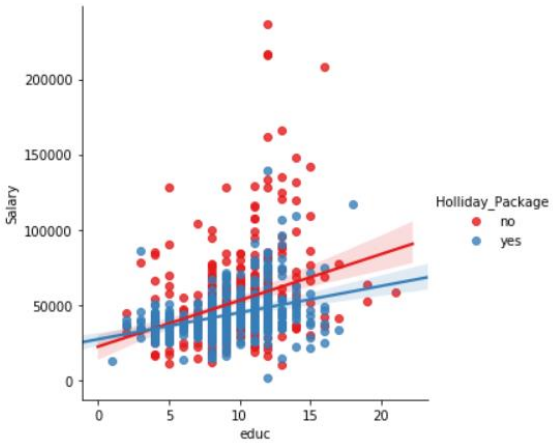

educ
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
no
yes
Holliday_Package

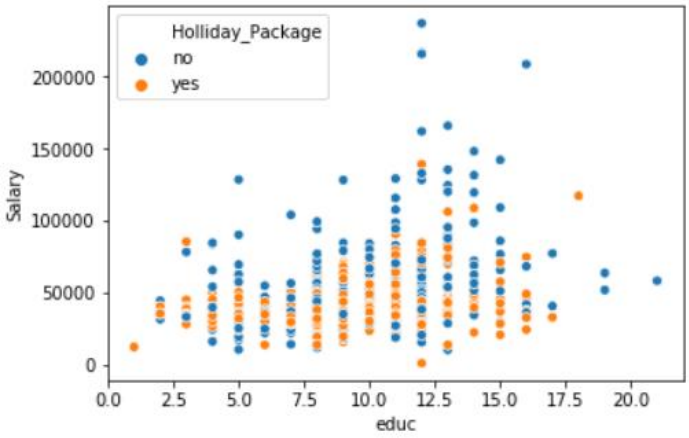

no_young_children
3.0
2.5
2.0
1.5
1.0
0.5
0.0
no
yes
Holliday_Package

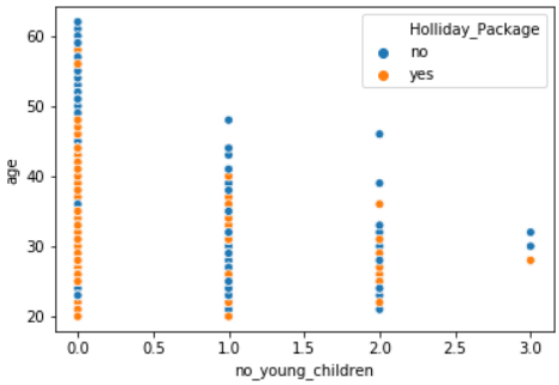

no_older_children
6
5
4
3
2
1
0
no
yes
Holliday_Package

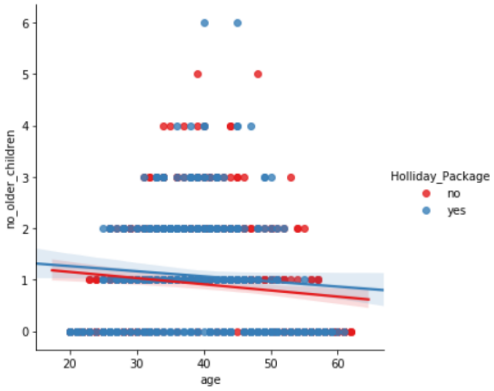








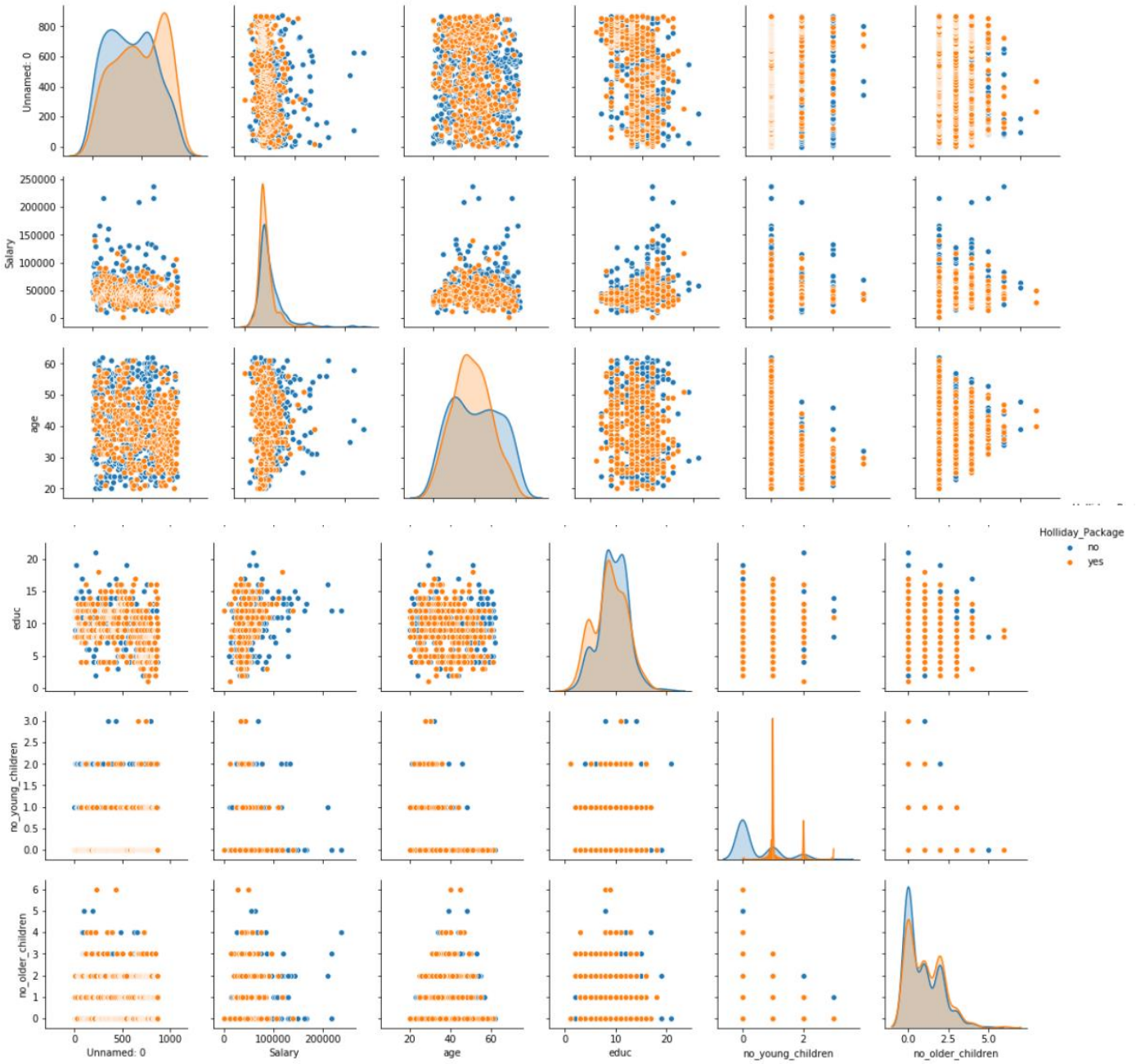




Holliday_Package	Unnamed: 0	Salary	age	educ	no_young_children	no_older_children	foreign
no	471	471	471	471	471	471	471
yes	401	401	401	401	401	401	401

Pair Plot

Pair plot visualizes given data to find the relationship between them where the variables can be continuous or categorical.



If we look at the distribution of record for holliday package opted and not opted , the opted part being interest to us is under representation of 401 as compared to not opted holliday package of 471. This kind of skew in the number of record we have in each class makes our overall accuracy score less reliable.

I we now observe our pair plot, What we see in the diagonal we see the estimates of density for each attributes. If we see the salary attribute both our classes are overlapping each other. Such attribute are unable to distinguish between the classes. Hence, they are not good attributes from classification point of view. Same is the case for age, no_older childrens and educ .

Skewness of every attribute

```
Unnamed: 0          0.000000
Salary              3.103216
age                 0.146412
educ               -0.045501
no_young_children   1.946515
no_older_children   0.953951
dtype: float64
```

The data is not properly skewed.

Logistic Regression

```
0.5305343511450382
[[133   9]
 [114   6]]
              precision    recall  f1-score   support

           0       0.54      0.94      0.68       142
           1       0.40      0.05      0.09       120

    accuracy                           0.53       262
   macro avg       0.47      0.49      0.39       262
weighted avg       0.48      0.53      0.41       262
```

0.530 - is overall accuracy . This is not reliable metrics as the classes are skewed.

1st row is for Not opted Holliday package class

2nd row is opted Holliday package class.

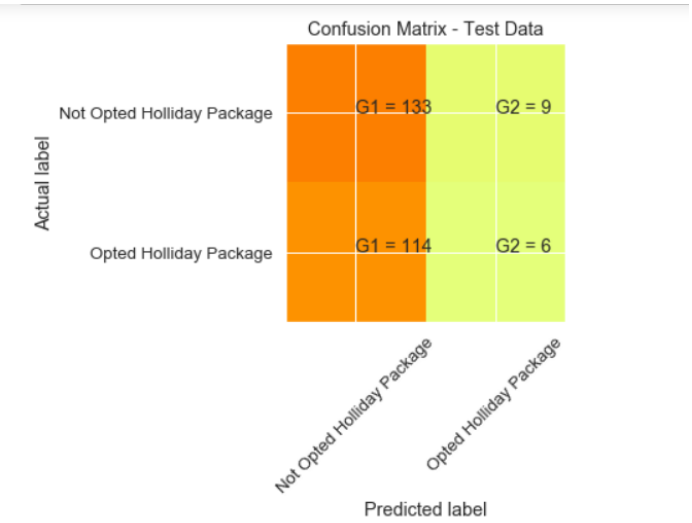
Similarly,

1st Column is for Not opted Holliday package class

2nd Column is for opted Holliday package class.

The score F1 is directly related to precision recall. (0- Not opted Holliday package class , 1- opted Holliday package class.)

Confusion matrix



G2- is true positive which shows 6 peoples were predicted correctly opted for Holliday package

G1- True negative 133

G2 - False positive (9)

G1 – False negative (114)

Performance Matrix

Training Data and Test Data Confusion Matrix Comparison

Confusion Matrix is a square matrix, which in the ideal case, its main diagonal must be valued and other sides must be none.

| TP | FP |

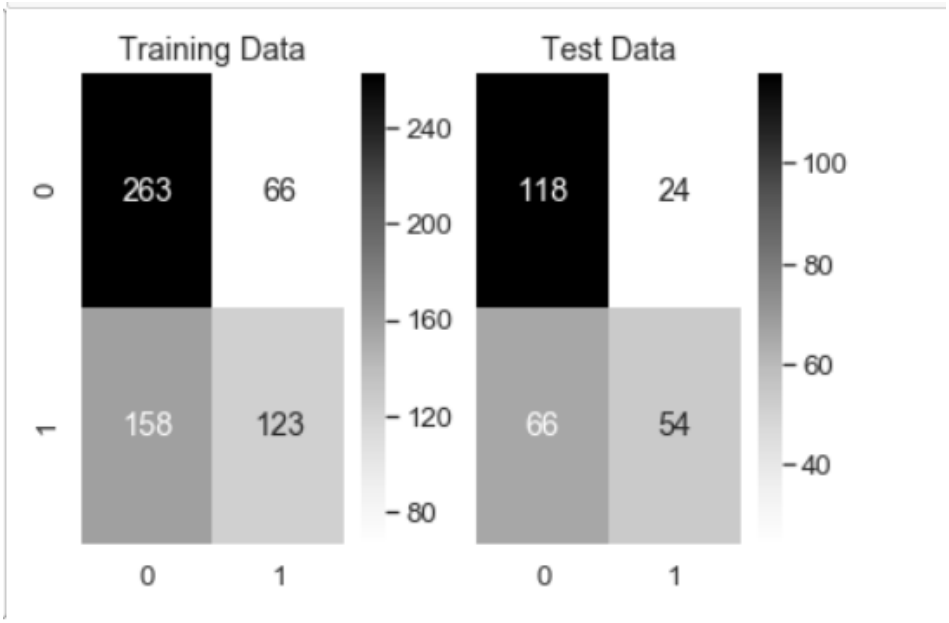
| FN | TN |

TP – True positive (correct data)

FP-False positive (data which is incorrect however still predicted as correct)

FN – False negative (Data which is correct however still predicted incorrect)

TN – True negative (Incorrect data)



Training Data and Test Data Classification Report Comparison

Classification Report of the training data:

	precision	recall	f1-score	support
0	0.62	0.80	0.70	329
1	0.65	0.44	0.52	281
accuracy			0.63	610
macro avg	0.64	0.62	0.61	610
weighted avg	0.64	0.63	0.62	610

Classification Report of the test data:

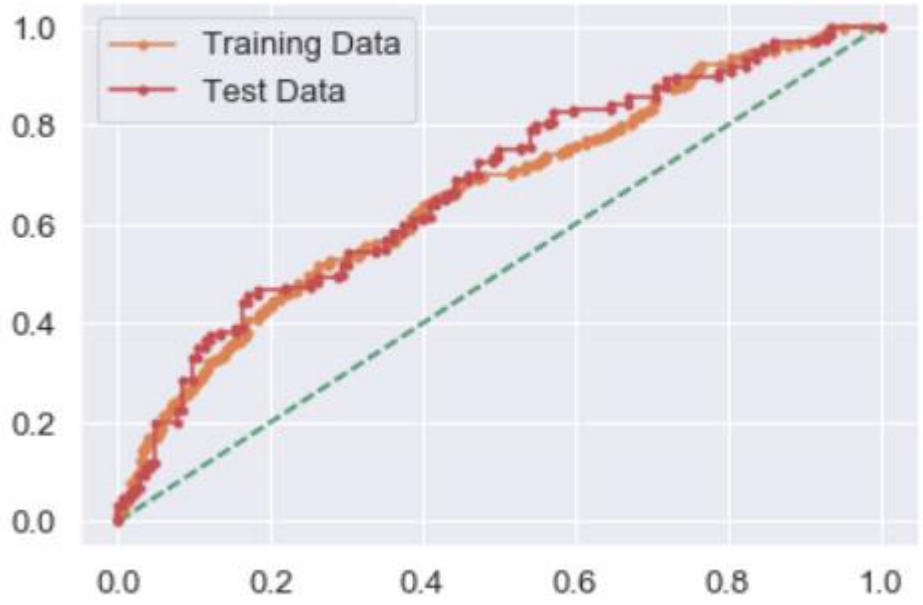
	precision	recall	f1-score	support
0	0.64	0.83	0.72	142
1	0.69	0.45	0.55	120
accuracy			0.66	262
macro avg	0.67	0.64	0.63	262
weighted avg	0.66	0.66	0.64	262

From the above we can conclude that performance metrix for test and train data is is almost same .

There is only little difference for f1-score , precision , recall for test and train data.

AUC for Training data and Test data

```
AUC for the Training Data: 0.661
AUC for the Test Data: 0.675
```



The model accuracy on the training as well as the test set is about 70%, which is roughly the same proportion as the class 0 observations in the dataset. This model is affected by a class imbalance problem. Since we only have 872 observations, if re-build the same LDA model with more number of data points, an even better model could be built.

Classification report of the default and custom cut-off test data

```
Classification Report of the default cut-off test data:

              precision    recall  f1-score   support

           0       0.64      0.83      0.72       142
           1       0.69      0.45      0.55       120

    accuracy                           0.66       262
   macro avg       0.67      0.64      0.63       262
weighted avg       0.66      0.66      0.64       262



Classification Report of the custom cut-off test data:

              precision    recall  f1-score   support

           0       1.00      0.04      0.07       142
           1       0.47      1.00      0.64       120

    accuracy                           0.48       262
   macro avg       0.73      0.52      0.35       262
weighted avg       0.76      0.48      0.33       262
```

Inference :

For both the models i.e. Logistic regression and linear discriminant analysis AUC, recall , precision , accuracy are almost same , hence both the models are best.